Valós idejű szövegosztályozás a Wikipédia szolgálatában

Solt Illés[1], Héder Mihály[2], Tikk Domonkos[1,3]

[1] Budapesti Műszaki és Gazd.tud. Egyetem, Távközl. és Médiainf. Tansz. (TMIT), H-1117 Budapest, Magyar Tud. krt. 2, e-mail: {solt,tikk}@tmit.bme.hu
[2] MTA SZTAKI, Internet Technológiák és Alkalmazások Központ (ITAK), H-1132 Budapest, XIII. Victor Hugo u. 18–22, e-mail: {mihaly.heder}@sztaki.hu
[3] Humboldt-Universität zu Berlin, Knowledge Management in Bioinformatics (WBI), D-10099 Berlin, Unter den Linden 6, e-mail: {tikk}@informatik.hu-berlin.de

A szöveges, azaz humán felhasználásra szánt információk hozzáférhetőségén javíthat, ha a szövegek több aspektus mentén is lekérdezhetőek. Ilyen rendszerre példa a Wikipédia közösségi enciklopédia, melyben az információ egysége a szócikk, melyek nem csak kereszthivatkozások mentén, hanem egy kategóriarendszer mentén is böngészhetőek. A szócikkek kategóriákba sorolásának koherenciája és a kategóriarendszer megválasztása határozza meg, hogy mennyivel könnyebben férhetők hozzá az összetartozó információk. Mind a kategóriarendszer kialakítását, mind a kategóriákba sorolást önkéntes szerkesztők végzik, akik nem feltétlenül ismerik a kategóriarendszert, melynek akár részleges feltárása is időigényes lehet, így nem várható el a szerkesztőtől. A szerkesztők munkája támogatásának kézenfekvő módja egy, a szócikk tartalma alapján kategóriaajánlatokat tevő rendszer. Itt bemutatunk egy elosztott, gyors válaszidejű, széleskörűen integrálható szövegosztályozó rendszert. Rendszerünk alkalmazhatóságát a magyar nyelvű Wikipédiába integrálható „okos” szerkesztővel demonstráljuk.

A szövegosztályozás, dokumentumok kategóriákba sorolása a természetes nyelvek feldolgozásának talán legkiforrottabb területe [1]. A szövegosztályozást a legtöbb módszer az alábbi lépésekben valósítja meg:

0. Nyers szöveggé alakítás (dokumentum → szöveg)
1. Nyelvi feldolgozás (szöveg → szófolyam): szavakra bontás, szótövezés, zajszavak eltávolítása;
2. Indexelés (szófolyam → egész vektor): egyedi szavak előfordulásainak összeszámlálása, a korpuszban túl gyakori vagy túl ritka szavak eltávolítása;
3. Súlyozás (egész vektor → valós vektor): a szavak dokumentumra vonatkozó fontosságának meghatározása,
4. Predikció (valós vektor → súlyozott kategóriák): betanított/felépített osztályozómodell alkalmazása.

Munkánk újdonsága nem a terület előremozdításában áll, hanem a szokásos offline, csővezeték jellegű feldolgozástól eltérő, közel valós idejű működésben. Tehát a bemutatásra kerülő rendszer a fenti lépéseket nem egy egész dokumentumgyűjteményre, hanem az egyes dokumentumokra külön végzi el, a válaszidőt előtérbe helyezve az átlagos feldolgozási idővel szemben. A kategóriajavaslatok

mellett a rendszer evidenciát is szolgáltat a döntésre a dokumentum adott kategóriára releváns szavainak kiemelésével.

A legtöbb fent vázolt technológiai lépés elvégzésére számos szabad szoftver (pl. NLTK, Snowball, Weka) és üzleti programcsomag található (pl. SPSS Textmining). Az itt bemutatásra kerülő megvalósításban[1] a nyelvi előfeldolgozást és indexelést az Apache Lucene[2], a súlyozást az Apache Mahout[3], az osztályozást pedig a HITEC osztályozó[4] [2] végzi. Az osztályozó választásakor a döntő szempont a HITEC mellett az volt, hogy támogatja a hierarchikus kategóriarendszereket, mint amilyen a Wikipédiáé is. A szócikkek nyers szöveggé alakítását a Devijver-féle elemző[5] módosított változata végzi.

Az osztályozó szolgáltatás a könnyű integrálhatóság érdekében HTTP REST felületen keresztül érhető el, a kimeneti formátumok között szerepel olvasható HTML és gépi feldolgozásra szánt XML. Az osztályozó példányok számának növelésével érhető el a rendszer horizontális skálázása, amely várhatóan elengedhetetlen Wikipédia méretű alkalmazás esetén.

A rendszer válaszideje 10 kB méretű (hosszú) szócikkekre 150 ms körüli, mely közel egyenlő arányban oszlik meg az előfeldolgozás (1–3) és a predikció (4) között. Ez a válaszidő ergonómiai szempontból nyilvánvalóan megfelel egy hálózati szolgáltatással szemben támasztott követelményeknek.

Az automatikus és azonnali kategóriajavaslatok felkínálása csak egy módja a Wikipédia-szerkesztők támogatásának. Folyamatban van az itt vázolt rendszer kiegészítése, mely a hasonló szócikkek, azok kategóriái, valamint a kategóriák jellemző szócikkeinek felkínálásával segíti a szerkesztőket a kategóriarendszerben való eligazodásban és így a jobb minőségű kategóriarendszer kialakításában.

Hivatkozások

1. Sebastiani, F.: Machine learning in automated text categorization. *ACM Computing Surveys* 2002; 34(1):1–47.
2. Tikk D., Biró Gy., Törcsvári A.: A hierarchical online classifier for patent categorization. *Emerging Technologies of Text Mining: Techniques and Applications* 2007; 244–67.

[1] `http://categorizer.tmit.bme.hu/trac/wiki/HITEC-java`
[2] `http://lucene.apache.org/`
[3] `http://mahout.apache.org/`
[4] `http://categorizer.tmit.bme.hu/trac/`
[5] `http://code.google.com/p/java-wikipedia-parser/`